doi:10.1093/nar/gkm794

MegaMotifBase: a database of structural motifs in protein families and superfamilies

Ganesan Pugalenthi[1], P. N. Suganthan[1], R. Sowdhamini[2,*] and Saikat Chakrabarti[3]

[1]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore, [2]National Centre for Biological Sciences, UAS-GKVK Campus, Bellary Road, Bangalore 560 065, India and [3]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received August 14, 2007; Revised September 16, 2007; Accepted September 17, 2007

ABSTRACT

Structural motifs are important for the integrity of a protein fold and can be employed to design and rationalize protein engineering and folding experiments. Such conserved segments represent the conserved core of a family or superfamily and can be crucial for the recognition of potential new members in sequence and structure databases. We present a database, MegaMotifBase, that compiles a set of important structural segments or motifs for protein structures. Motifs are recognized on the basis of both sequence conservation and preservation of important structural features such as amino acid preference, solvent accessibility, secondary structural content, hydrogen-bonding pattern and residue packing. This database provides 3D orientation patterns of the identified motifs in terms of inter-motif distances and torsion angles. Important applications of structural motifs are also provided in several crucial areas such as similar sequence and structure search, multiple sequence alignment and homology modeling. MegaMotifBase can be a useful resource to gain knowledge about structure and functional relationship of proteins. The database can be accessed from the URL http://caps.ncbs.res.in/MegaMotifbase/index.html

INTRODUCTION

Previous studies have pointed out that short segments of sequence and/or structural elements are required for retention of fold and function of a protein (1,2). Sequence-based representations, however, are only an approximation to the underlying structural and functional information. Therefore, motifs identified at 3D structure level provide significant and reliable information. We had earlier identified such structurally invariant segments through careful manual intervention for superfamilies where proteins are distantly related but retain similar fold and biological functions (3,4).

Here, we present a database, MegaMotifBase, which provides a set of important structural segments or motifs for protein structures related at family or superfamily level on the basis of conservation of both sequence and structural features. Motifs among structurally aligned proteins are recognized by the conservation of amino acid preference and solvent accessibility and are examined for the conservation of important structural features like secondary structural content, hydrogen-bonding pattern and residue packing (3–5). These motifs may form the common structural core by maintaining a particular spatial pattern, when compared across different proteins belonging to the same family or superfamily. Such motifs can also be employed to design and rationalize protein engineering and folding experiments. MegaMotifBase provides a comprehensive compilation of structural motifs identified through a completely automated method for large number of families (1032) and superfamilies (1194) of proteins, in contrast to earlier efforts (3,4,6), which were limited to poor coverage and extensive manual supervision. This database can be accessed from the URL http://caps.ncbs.res.in/MegaMotifbase/index.html.

KEY FEATURES OF THE DATABASE

- Identification and collection of important conserved structural segments that are crucial for the integrity of the fold and can be projected as the minimum structural requirements for a new member to be considered as part of a pre-existing family or superfamily. It is also possible to use simple sequence conservation to recognize motifs.

*To whom correspondence should be addressed. Tel: +91-80-23666250; Fax: +91-80-3636662; Email: mini@ncbs.res.in
Correspondence may also be addressed to Saikat Chakrabarti. Tel: 301-594-6474; Email: chakraba@mail.nlm.nih.gov

- Interactive 3D views of the motifs on the superposed structures are displayed for better understanding and visualization.
- Spatial orientations of the motifs, in terms of inter-motif distances and torsion angles, are provided. This enables the users to analyze the structural variations that are felt even at conserved core of the fold owing to poor sequence identity and evolutionary pressures.
- Options are provided for scanning multiple structural motifs along with their spatial orientation in a given query protein structure and to scan multiple motifs in a query structure against the entire structural motif database. This could be very useful in protein classification and assignment of family or superfamily relationship to newly solved protein structures with unknown function.
- Options are also provided to search for similar sequences by a multimotif-based database scanning procedure called SCANMOT (7). This scanning option provides an opportunity to identify distantly related sequences for each family or superfamily. The specificity of the search engine is increased by utilizing the inter-motif spacing and pairwise global alignment of the query and hits.
- The current version of the database provides options to align similar sequence with the query protein structure(s). It allows the user to obtain a control over the alignment by providing sequence–structure motif regions as input to the alignment program to achieve a more structurally relevant and functionally useful alignment of protein sequences. The alignment algorithm employs local conserved regions of the sequences to be fixed, and aligns the rest based on normal progressive alignment. The chances of global misalignment are thereby reduced and the possibility of obtaining overall better alignment is increased (8).
- The database also allows users to build 3D models of similar protein sequences of unknown structure.
- The entire database of motifs and the alignments can be downloaded as a flat file for further use and analyses (Figure 1).

CONTENTS OF THE DATABASE

MegaMotifBase compiles structural motifs at different levels of protein classification strata.

Structural motifs at the family level

We have collected 1032 structural alignments of protein domains that are related at the family levels from the HOMSTRAD (9) database. Motifs among structurally aligned proteins were recognized by the conservation of amino acid preference and solvent accessibility and examined for the conservation of other important structural features like secondary structural content, hydrogen-bonding pattern and residue packing. Sequentially conserved regions were identified from the multiple alignments by examining the nature of amino acid exchanges using a standard 20 × 20 substitution matrix (10). Solvent accessibility was measured using the PSA program from JOY4.0 suite (11). SSTRUC and HBOND programs, that are also part of JOY4.0 suite, were used to identify secondary structural positions and hydrogen bonds, respectively. Residue packing has been measured in terms of Ooi number (12) that provides the number of residues surrounding each C^{α} atom of residues in a protein. Higher Ooi numbers correspond to better residue packing and suggest that the residue is in a well-packed environment.

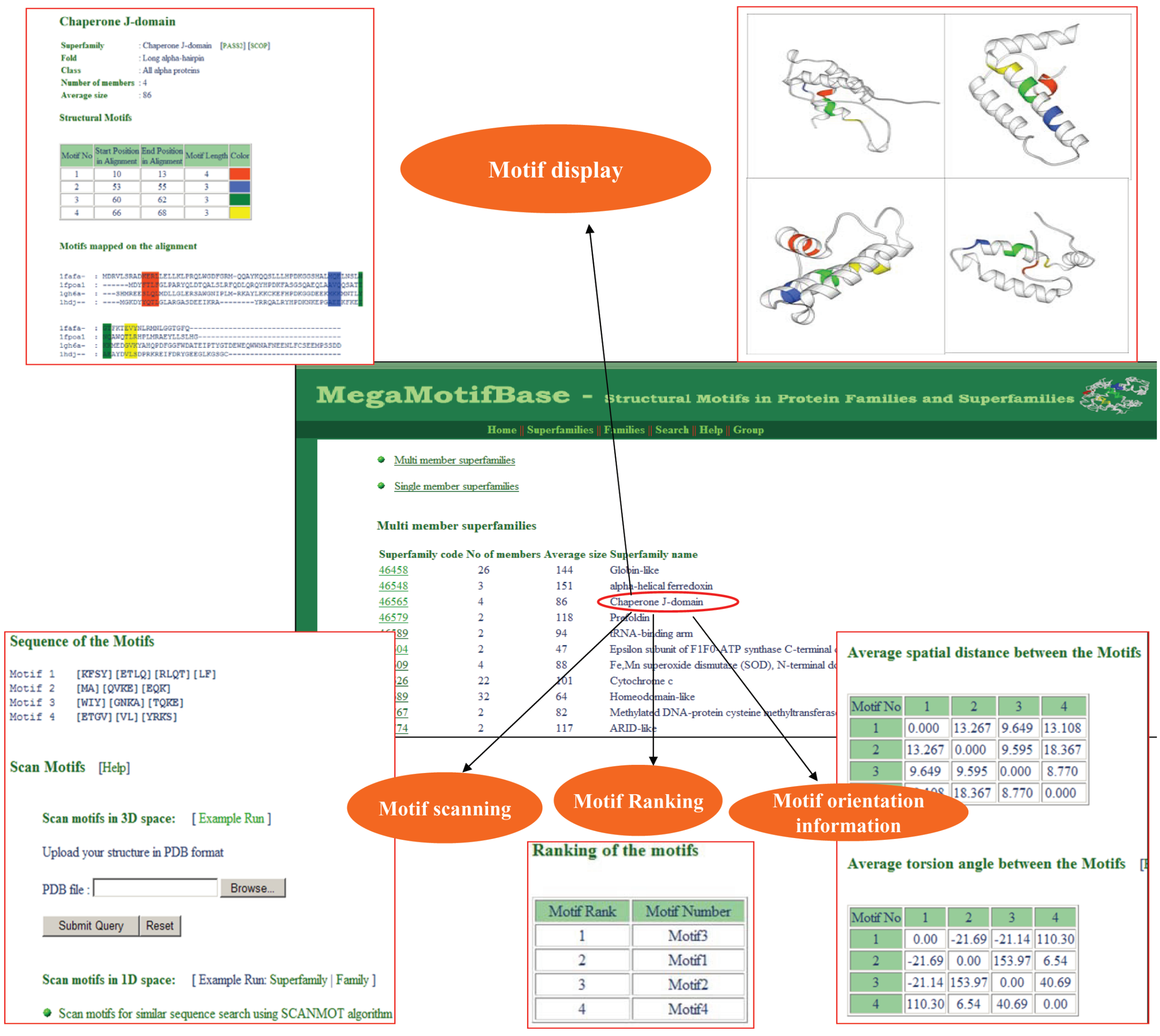
A structural feature is considered conserved at an alignment position if it is present in all or all but one member within the alignment. We found this condition was more practical for families with poor structural representation. The structural motifs identified are mapped on the alignment using different color code and often represent the conserved core of the family. Ranking of the motifs is performed considering the extent of conservation of the structural feature. An idea of the 3D orientation pattern of the structural motifs is provided via graphic displays and spatial orientation matrices.

Structural motifs at the superfamily level

Structural motifs for multimember superfamilies. The superfamily is a hierarchical classification that contains proteins of different families having similar structure and function. These proteins might have very low sequence identities but retain the same fold through well-conserved secondary structural parts. Therefore, identification of structural motifs for superfamilies is even more valuable since the evolutionary divergence makes it impossible to derive conserved sequence or structural segments simply by residue conservation. We identified structural motifs for all the multimember superfamilies (628) available in the latest PASS2 and SCOP (version 1.63) databases (13,14) following the same protocol described above to identify motifs for proteins related at the family level.

Structural motifs for single member superfamilies. A majority of the entries in protein structural databank are single member superfamilies for which it is hard to derive structural motifs due to the paucity of structural homologues. Important conserved segments for these 566 superfamilies (PASS2 database, (13)) have been identified and compiled into the MegaMotifBase. Conserved regions, recognized by permitted amino acid exchanges, were mapped onto the structure and content of various structural features (solvent accessibility, secondary structure content, hydrogen bonding and residue packing) were examined. Only the conserved segments with high structural feature content were projected as sequence-structural templates for the particular superfamily member. Interactive 3D displays of the templates in 3D structure [in Chime® and RASMOL (15)] were provided for better understanding and visualization. A static image of the 3D structure is provided using MOLSCRIPT (16).

We also provide the application of sequence–structural templates in three different areas: multimotif-based



Chaperone J-domain

Superfamily : Chaperone J-domain [PASS2] [SCOP]
Fold : Long alpha-hairpin
Class : All alpha proteins
Number of members : 4
Average size : 86

Structural Motifs

Motif No	Start Position in Alignment	End Position in Alignment	Motif Length	Color
1	10	13	4	🟥
2	53	55	3	🟦
3	60	62	3	🟩
4	66	68	3	🟨

Motifs mapped on the alignment

```
1fafa-  : MDRVLSRADKERLLELLKLPRQLWGDFGRM-QQAYKQQSLLLHPDKGGSHALMQELNSLW
1fpoa1  : ------MDYFTLFGLPARYQLDTQALSLRFQDLQRQYHPDKFASGSQAEQLAAVQQSATI
1gh6a-  : ---SHMREESLQLMDLLGLERSAWGNIPLM-RKAYLKKCKEFHPDKGGDEEKMKKMNTLY
1hdj--  : ----MGKDYYQTLGLARGASDEEIKRA--------YRRQALRYHPDKNKEPGAEEKFKEI

1fafa-  : GTFKTEVYNLRMNLGGTGFQ-----------------------------------
1fpoa1  : NQAWQTLRHPLMRAEYLLSLHG---------------------------------
1gh6a-  : KKMEDGVKYAHQPDFGGFWDATEIPTYGTDEWEQWWNAFNEENLFCSEEMPSSDD
1hdj--  : AKAYDVLSDPRKREIFDRYGEEGLKGSGC--------------------------
```

MegaMotifBase - Structural Motifs in Protein Families and Superfamilies

Home || Superfamilies || Families || Search || Help || Group

- Multi member superfamilies
- Single member superfamilies

Multi member superfamilies

Superfamily code	No of members	Average size	Superfamily name
46458	26	144	Globin-like
46548	3	151	alpha-helical ferredoxin
46565	4	86	Chaperone J-domain
46579	2	118	Prefoldin
589	2	94	tRNA-binding arm
04	2	47	Epsilon subunit of F1F0-ATP synthase C-terminal d
09	4	88	Fe,Mn superoxide dismutase (SOD), N-terminal d
26	22	101	Cytochrome c
89	32	64	Homeodomain-like
67	2	82	Methylated DNA-protein cysteine methyltransferas
74	2	117	ARID-like

Sequence of the Motifs

```
Motif 1    [KFSY][ETLQ][RLQT][LF]
Motif 2    [MA][QVKE][EQK]
Motif 3    [WIY][GNKA][TQKE]
Motif 4    [ETGV][VL][YRKS]
```

Scan Motifs [Help]

Scan motifs in 3D space: [Example Run]

Upload your structure in PDB format

PDB file : ______ Browse...

Submit Query | Reset

Scan motifs in 1D space: [Example Run: Superfamily | Family]

- Scan motifs for similar sequence search using SCANMOT algorithm

Ranking of the motifs

Motif Rank	Motif Number
1	Motif3
2	Motif1
3	Motif2
4	Motif4

Average spatial distance between the Motifs

Motif No	1	2	3	4
1	0.000	13.267	9.649	13.108
2	13.267	0.000	9.595	18.367
3	9.649	9.595	0.000	8.770
[illegible]	[illegible]	18.367	8.770	0.000

Average torsion angle between the Motifs [I

Motif No	1	2	3	4
1	0.00	-21.69	-21.14	110.30
2	-21.69	0.00	153.97	6.54
3	-21.14	153.97	0.00	40.69
4	110.30	6.54	40.69	0.00

Figure 1. A snapshot image capturing the key features of the MegaMotifBase database for an example superfamily.

sequence search, multiple sequence alignment and homology modeling. In each case, the inclusion of the sequence-structural templates can give rise to sensitive and accurate results. This emphasizes the need for inclusion of singletons to provide added value to the recognition of additional members, comparative modeling and in designing experiments.

APPLICATIONS

The availability of structural motifs is useful since these conserved patterns form the common core by maintaining a particular spatial orientation pattern. These motifs can also assist in the identification of new potential members of an existing family and/or superfamily. Scanning of multiple structural motifs, along with their spatial orientation in a given query protein structure, could be very useful in protein structural classification. In MegaMotifBase database, we also provide the application of motifs in three other crucial areas: motif-based similar sequence search, multiple sequence alignment and in homology modeling. In each case, the inclusion of the sequence–structural motifs can give rise to sensitive and accurate results.

ACKNOWLEDGEMENTS

G.P. and P.N.S. acknowledge the financial support offered by A*Star (Agency for Science, Technology and Research). R.S. acknowledges Wellcome Trust, UK and National Centre for Biological Sciences (TIFR) for financial and infrastructural support. S.C. acknowledges Intramural Research Program of the National Library of Medicine at NIH/DHHS. Funding to pay the Open Access publication charges for this article was provided by Wellcome Trust, U.K.

Conflict of interest statement. None declared.

REFERENCES

1. Farber,G.K. and Petsko,G.A. (1990) The evolution of alpha/beta barrel enzymes. *Trends Biochem. Sci.*, **15**, 228–234.
2. Kannan,N., Selvaraj,S., Gromiha,M.M. and Vishveshwara,S. (2001) Clusters in alpha/beta barrel proteins: implications for protein structure, function, and folding: a graph theoretical approach. *Proteins*, **43**, 103–112.
3. Chakrabarti,S., Venkatramanan,K. and Sowdhamini,R. (2003) SMoS: a database of structural motifs of protein superfamilies. *Protein Eng.*, **16**, 791–793.
4. Chakrabarti,S. and Sowdhamini,R. (2004) Regions of minimal structural variation among members of protein domain superfamilies: application to remote homology detection and modeling using distant relationships. *FEBS Lett.*, **569**, 31–36.
5. Pugalenthi,G., Suganthan,P.N., Sowdhamini,R. and Chakrabarti,S. (2007) SMotif: a server for structural motifs in proteins. *Bioinformatics*, **23**, 637–638.
6. Chakrabarti,S., Manohari,G., Pugalenthi,G. and Sowdhamini,R. (2006) SSToSS—sequence-structural templates of single-member superfamilies. *In Silico Biol.*, **6**, 311–319.
7. Chakrabarti,S., Anand,A.P., Bhardwaj,N., Pugalenthi,G. and Sowdhamini,R. (2005) SCANMOT: searching for similar sequences using a simultaneous scan of multiple sequence motifs. *Nucleic Acids Res.*, **33**, W274–W276.
8. Chakrabarti,S., Bhardwaj,N., Anand,P.A. and Sowdhamini,R. (2004) Improvement of alignment accuracy utilizing sequentially conserved motifs. *BMC Bioinformatics*, **5**, 167–178.
9. Mizuguchi,K., Deane,C.M., Blundell,T.L. and Overington,J.P. (1998) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.*, **7**, 2469–2471.
10. Johnson,M.S. and Overington,J.P. (1993) A structural basis for sequence comparisons. An evaluation of scoring methodologies. *J. Mol. Biol.*, **233**, 716–738.
11. Mizuguchi,K., Deane,C.M., Blundell,T.L., Johnson,M.S. and Overington,J.P. (1998) JOY: protein sequence-structure representation and analysis. *Bioinformatics*, **14**, 617–623.
12. Nishikawa,K. and Ooi,T.J. (1986) Radial locations of amino acid residues in a globular protein: correlation with the sequence. *J. Biochem.*, **100**, 1043–1047.
13. Bhaduri,A., Pugalenthi,G. and Sowdhamini,R. (2004) PASS2: an automated database of protein alignments organised as structural superfamilies. *BMC Bioinformatics*, **5**, 35–41.
14. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
15. Sayle,A. and Milner-White,E. J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374–376.
16. Kraulis,P.J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.*, **24**, 946–950.